Educational Researcher

http://er.aera.net

Common Structural Design Features of Rubrics May Represent a Threat to Validity

Stephen Mark Humphry and Sandra Allison Heldsinger

EDUCATIONAL RESEARCHER 2014 43: 253 originally published online 27 June 2014
DOI: 10.3102/0013189X14542154

The online version of this article can be found at:
http://edr.sagepub.com/content/43/5/253

Published on behalf of

American Educational Research Association

and

SAGE

http://www.sagepublications.com

Additional services and information for *Educational Researcher* can be found at:

Email Alerts: http://er.aera.net/alerts

Subscriptions: http://er.aera.net/subscriptions

Reprints: http://www.aera.net/reprints

Permissions: http://www.aera.net/permissions

>> Version of Record - Jul 14, 2014

OnlineFirst Version of Record - Jun 27, 2014

What is This?

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

Common Structural Design Features of Rubrics May Represent a Threat to Validity

Stephen Mark Humphry[1] and Sandra Allison Heldsinger[1]

Rubrics for assessing student performance are often seen as providing rich information about complex skills. Despite their widespread usage, however, little empirical research has focused on whether it is possible for rubrics to validly meet their intended purposes. The authors examine a rubric used to assess students' writing in a large-scale testing program. They present empirical evidence for the existence of a potentially widespread threat to the validity of rubric assessments that arose due to design features. In this research, an iterative tryout-redesign-tryout approach was adopted. The research casts doubt on whether rubrics with structurally aligned categories can validly assess complex skills. A solution is proposed that involves rethinking the structural design of the rubric to mitigate the threat to validity. Broader implications are discussed.

Keywords: assessment; item response theory; mixed methods; performance assessment; teacher assessment; test theory/development; validity/reliability

At a time when advocacy for performance-based assessment was gaining momentum, Messick (1994) raised the question of whether rubrics validly meet the purposes of their usage, asking the following: "By what evidence can we be assured that the scoring criteria and rubrics used in holistic, primary trait, or analytic scoring of products or performances capture the fully functioning complex skill?" (p. 20). Nearly two decades later, even though the use of rubrics is now widespread across the globe, surprisingly little empirical research has been devoted to answering this question (Reddy & Andrade, 2010; Rezaei & Lovorn, 2010).

A number of important issues have been canvassed and discussed in the literature within different conceptual frameworks (e.g. Kohn, 2006). However, little experimental research is available to inform these discussions (Meier, Rich, & Cady, 2006). The present article aims to begin to address this issue by presenting and examining evidence of a potential threat to valid assessment and by presenting empirical evidence indicating the existence of specific issues identified in the literature, particularly by Sadler (2009).

On the basis of the empirical evidence, we will argue that the widely used matrix design of rubrics can create a threat to valid performance assessment. The threat arises because there is typically no underlying developmental or learning theory that justifies having the same number of qualitative gradations across criteria. The focus here is on *construct validity*, as this term is defined later. We aim to show that rethinking the structural design features of rubrics may avoid this specific threat to validity by allowing rubrics to more faithfully capture qualitative gradations of performance independently for each criterion.

Our intention is not to claim that resolving the threat to validity addresses other validity-related issues (e.g., whether the task and criteria are appropriate to assess a trait). However, we propose that resolving the validity threat opens the way for more productive research into a number of questions, such as to ascertain which and how many criteria should be used, whether the operational independence of criteria can be established, and the optimal number of qualitative gradations for each separate criterion. Resolving the threat to validity might also open the way to more productive research into whether raters make more valid assessments using rubrics than holistic judgments.

We present evidence in the article based on empirical research conducted in the context of the assessment of narrative writing in a large-scale, standardized testing program. The evidence indicates that the typical grid or matrix design of the rubric used in this context induced pronounced rating tendencies of a form that would usually be interpreted to indicate a halo effect. The term

[1]The University of Western Australia, Crawley, WA

Educational Researcher, Vol. 43 No. 5, pp. 253–263
DOI: 10.3102/0013189X14542154
© 2014 AERA. http://er.aera.net

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

halo effect refers to a strong tendency for ratings on separate items or criteria to reflect a general rater impression of a performance.

We adopted an iterative tryout-redesign-tryout approach (Ercikan & Roth, 2006) to investigate the source of the rating tendencies. As part of this approach, we developed a conceptual framework in which we propose that the instrument can play a decisive role in inducing rating tendencies. In particular, we propose that the matrix design of a rubric may create rating tendencies. We aim to show that restructuring the rubric to avoid a matrix design substantially reduced the threat to validity. The broader implications of the empirical findings are discussed.

The Typical Design Structure of Rubrics Used in Performance Assessment

Performance- and product-based assessments are seen as providing teachers with rich information about student competence, leading to positive consequences for teaching and learning (Darling-Hammond, 1994; Messick, 1994). The advocacy for performance-based assessment was in large part a reaction to multiple-choice tests, which were criticized for decontextualization and skill decomposition.

Performance assessment consists of two parts—a task and a set of scoring criteria or rubric (Perlman, 2003). A rubric "lists the criteria for a piece of work of what counts and articulates gradations of quality for each criterion" (Andrade, 2005, p. 27). Typically rubrics are presented as a grid in which each criterion has the same number of gradations of quality. One of the appeals of rubrics is that they are easily constructed and readily interpreted (Andrade, 2000). However, the ease of construction may come at a cost to the validity of assessments.

There is debate as to whether a rubric needs to be task-specific so that it applies to a single task or generic so that the same rubric can be applied to a number of different tasks (Popham, 1997; Wiliam, 2011). The debate emanates from the desire for rubrics to have broader applicability and thereby to help students generalize learning from one context to another. "Rubrics are often used by teachers to grade student work but many authors argue that they can serve another, more important, role as well: When used by students as part of a formative assessment of their works in progress, rubrics can teach as well as evaluate" (Reddy & Andrade, 2010, p. 437).

It seems rubrics have captured the imagination of a large proportion of the educational community, particularly in recent decades. Many educational resources include rubrics as a matter of course and numerous websites provide teachers with easy ways to generate rubrics. One such website, *Rubistar*, recorded traffic for the school year 2011-2012 of 1,493,317 unique visits, and 2,465, 985 visits overall (personal correspondence, Ault, May 17, 2012).

Rubrics are used in early childhood education and across subjects in the school years (Rezaei & Lovorn, 2010; Tierney & Simon, 2004). They are widely used across a range of disciplines in higher education (Sadler, 2009), and more recently have been used to evaluate teachers (Papay, 2012). Rubrics are perceived to cross the traditional divide of formative and summative assessments and are used as informal assessments in the classroom, as well as in many standardized assessment programs. Yet there is a dearth of empirical research on the quality of rubrics as assessment instruments and the research that is reported tends to be based on small-scale studies.

Evaluation of Rubrics

Reddy and Andrade (2010) conducted a critical review of the empirical research on the use of rubrics at the postsecondary level and found that the large majority of studies did not describe the process of development of rubrics to establish their quality. Their review identified four areas most in need of attention from the scholarly community: rigorous research methodologies, geographical focus, validity and reliability, and the promotion of learning. Of particular relevance here is the recommendation for further research into the validity and reliability of assessments using rubrics. The authors found that "some studies mention having conducted pilot and reliability tests prior to the implementation of rubrics, however very few report the results" (Reddy & Andrade, 2010, p. 446). The authors go on to make a recommendation: "Future studies should report how the validity of a rubric was established, and the scoring reliability, including rater training and its contribution toward achieving inter-rater reliability, and perhaps even the correlation between rubric-referenced scores and other measures of performance" (Reddy & Andrade, 2010, p. 446).

Several studies have found rubrics provide reliable judgments (Bresciani et al., 2009; Jonsson & Svingby, 2007; Silvestri & Oescher, 2006; Wald et al., 2012). Moskal and Leydens (2000) state that rubrics address concerns of subjectivity and that, by formalizing the criteria for scoring a student product or performance, they can reduce variations between raters.

However, some educators have questioned the assumption that the use of rubrics increases interrater reliability and validity, and the overall accuracy and quality of assessment (Delandshere & Petrosky, 1998; Wilson, 2006, as cited in Rezaei & Lovorn, 2010). Rezaei and Lovorn (2010) reported a study in which participants were asked to grade one of the two samples of writing, assuming it was written by a graduate student, once using a rubric and once without a rubric. Their results showed that the raters were significantly influenced by mechanical characteristics of the students' writing rather than the content, even when they used the rubric. This led them to ask: "if a rubric like the one used in this project, which was designed by a group of professors in a college of education, is shown to be unreliable, then what does this say about the thousands of rubrics being used every day in schools?" (Rezaei & Lovorn, 2010, p. 29). In addition to the question of the reliability of rubric assessments, there remains the concern raised by Messick (1994, p. 14): "The portrayal of performance assessment as *authentic* and *direct* has all the earmarks of a validity claim but with little or no evidential grounding."

Theoretical Background

Validity

In his authoritative accounts of validity, Messick (1989, 1994) described two major threats to construct validity, namely "construct-underrepresentation (which jeopardizes authenticity) and construct-irrelevant variance (which jeopardizes directness)" (Messick, 1994, p. 14). Although Borsboom, Mellenbergh, and van Heerden (2004) are critical of Messick's validity theory, they

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

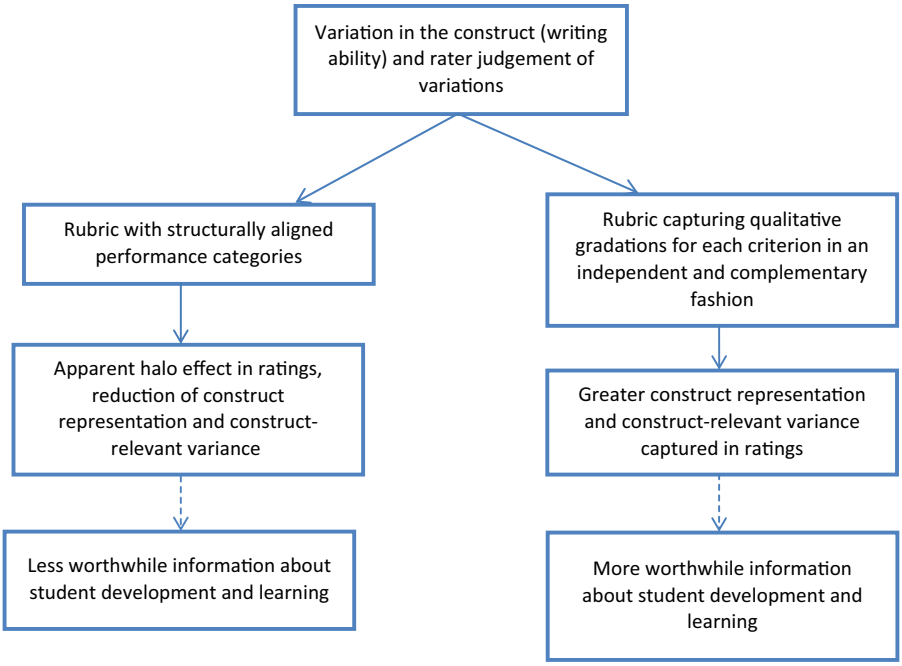

FIGURE 1. *Schematic depiction of the role of an assessment rubric in creating an apparent halo effect*

clearly argue that construct-relevant variance is essential to validity, stating that for measurement to be valid, it is necessary that "variations in the attribute [or construct] causally produce variations in the outcomes of the measurement procedure" (p. 1061). Thus, according to either of these prominent perspectives on validity, threats to construct representation and construct-relevant variance constitute threats to construct validity.

In this article, we adopt Borsboom et al.'s (2004) central criterion for validity, that variations in the attribute must produce variations in measurement outcomes. These authors argue for the parsimonious definition: Construct validity is about whether a test measures what it is designed to measure. We hypothesize that the structural alignment of rubric categories seriously limits the degree to which variations among students' writing abilities can result in corresponding variations among test scores, as required to obtain a valid assessment of writing according to their central criterion and definition.

The Halo Effect and Rating Tendencies

If the halo effect occurs because judgments are strongly influenced by a global impression, it can clearly prevent construct-relevant variance. The halo effect has been extensively studied along with other rater effects such as leniency, central tendency, and restriction of range (Myford & Wolfe, 2004). Thorndike (1920) coined the term *halo*, when he reviewed findings from a 1915 study where it appeared "that the estimates of the same man in a number of different traits such as intelligence, industry, technical skill, reliability, etc., etc., were very highly correlated and very evenly correlated" (Thorndike, 1920, p. 25). Thorndike (1920) concluded the "ratings were apparently affected by a marked tendency to think of the person in general as rather good or rather inferior and to color the judgments of the qualities by this general feeling" (p. 25). Similarly, such a tendency prevents raters from judging individual differences in a construct related to separate aspects of performances, thus preventing construct-relevant variation in scores.

The majority of conceptual definitions of the halo effect can be grouped into three categories based on the stated or implied cause: (i) the influence of a general impression of ratees, (ii) the influence of a salient characteristic of ratees, and (iii) inadequate discrimination of ratees by raters (Fisicaro & Lance, 1990; Fisicaro & Vance, 1994).

Fisicaro and Lance (1990) propose causal models of these three categories of halo effects based on the three categories of definition. Their models incorporate only rater and ratee influences on ratings, and they do not take into account the role of assessment instruments. From this broader perspective, therefore, the models indicate a broader lack of attention to the possibility that design features of assessment instruments may result in unjustifiably high correlations of ratings on separate assessment criteria.

Conceptual Framework

The Importance of Structural Rubric Design Features to Construct Validity

A typical rubric has a matrix design in which each criterion has the same number of gradations of quality. The following question arises: Why should there be precisely the same number of gradations of quality for each criterion? We propose that this a priori alignment may induce strong rating tendencies that would usually be interpreted as a halo effect. Furthermore, we propose that these tendencies (i) cannot be attributed solely to rater behavior and (ii) do not fit with any of the three categories of definitions of the halo effect stated above.

To describe the role played by structural design features, we developed the conceptual framework depicted in Figure 1. For the sake of simplicity, the framework describes two polar extremes in which a rubric either (i) induces strong rating

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

Table 1
Examples of Category Descriptors From the Stage 1 Rubric

Criterion	Category 1	Category 2
Form of writing	Demonstrates a beginning sense of story structure, for example opening may establish a sense of narrative.	Writes a story with a beginning and a complication. Two or more events in sequence. May attempt an ending.
Subject matter	Includes few ideas on conventional subject matter, which may lack internal consistency.	Has some internal consistency of ideas. Narrative is predictable. Ideas are few, may be disjointed, and are not elaborated.
Text organisation	Attempts sequencing although inconsistencies are apparent.	Writes a text with two or more connected ideas. For longer texts, overall coherence is not observable.

tendencies or (ii) facilitates complementary judgments based on separate criteria.

Our justification for the conceptual framework is as follows. If there are separate assessment criteria and the aim is to describe performance categories within each of the criteria, there is no a priori reason to expect that there should be the same number of qualitatively distinguishable performance levels in one criterion as in any other in the rubric. Typically, there is no underlying developmental or learning theory that justifies having precisely the same number of qualitatively distinguishable stages across multiple aspects of a construct. This makes it unlikely that the gradations of quality faithfully capture that which is observed in student performances for each criterion separately from other criteria. It is more likely instead that the same numbers of gradations of quality are chosen for convenience in constructing rubrics and for ease of marking than because equal numbers of gradations faithfully capture the distinguishable performance levels for separate criteria. That is, it is decided a priori that each criterion has the same number of gradations of quality rather than this decision having been based on theoretical or empirical grounds.

We propose that structurally aligned gradations of quality tend to induce rating tendencies and score patterns such as (1, 1, 1, …), (2, 2, 2, …), and so on. The structural alignment forces ratings to be artificially alike, thus limiting raters from capturing variation in separate aspects of the construct. In this sense, structural alignment precludes construct-relevant variance in the scores for any given student, thus failing to meet a basic criterion for validity articulated by Messick (1994) and Borsboom et al. (2004).

We do not mean to imply that criteria need to be wholly or even largely independent of each other. Criteria may be mutually related by virtue of their reference to the common construct. Nevertheless, given the aim of using a rubric, it is desirable for criteria to contain descriptions of performances free of obvious overlap or redundancy to allow raters to focus on distinctive and complementary aspects of students' performances, and to capture individual differences in each aspect within ratings.

Hypothesized Source of the Apparent Halo Effect

We hypothesize that the structural alignment of categories can produce an apparent halo effect for two reasons. The first reason is that structural alignment, where criteria have equal numbers of categories, may result in more or less categories for any given criterion than is optimal given the number of qualitative distinctions that raters can make. If there are too many categories, judges may have little choice but to make spurious distinctions either by defaulting to a pattern of common scoring (akin to a response set) or through recourse to a global judgment. (The issue is not the use of a global judgment per se but rather that repetition of a global judgment is contrary to the aim of analytic scoring.) If, on the other hand, there are too few categories, judges are prevented from making distinctions they are capable of making. In this case, again, ratings do not reflect variation in the quality of performances that raters can discern.

The second reason is that structural alignment can create a degree of unintended conceptual overlap and redundancy in the descriptions of gradations for some pairs of criteria as described by Sadler (2009, p. 169). To illustrate this, we use an actual example in Table 1, which shows an extract taken from the original rubric that provided the impetus for the research that we report in this article. The conceptual overlap between the descriptions in the criteria is evident as follows. If a student has provided a beginning and a complication as described in Category 2 of *Form of Writing*, the student has almost by definition provided a narrative that contains two or more related ideas, as described in Category 2, *Text Organisation.* A narrative that contains two or more consistent ideas will necessarily have demonstrated some internal consistency of ideas as described in Category 2 of the criteria *Text Organisation* and *Subject Matter*, respectively.

The design and structure of the rubric can therefore constrain raters to award the same, or highly similar, scores across the criteria. In practice, such a rating tendency is likely to be (mis) interpreted as a halo effect; that is, it is likely to be interpreted as "the tendency of a rater to allow overall impressions of an individual to influence the judgements of that person's performance along several quasi-independent dimensions of [performance]" (King, Hunter, & Schmidt, 1980, p. 507).

The conceptual framework was developed on the basis of the findings from a series of studies that investigated an apparent halo effect. The empirical research that will be presented indicates that when a rubric faithfully captures qualitative gradations independently for each criterion, the judgments more faithfully reflect construct-relevant variation that raters can discern within each criterion separately. The empirical findings are reported next.

Empirical Studies

Background

The research was carried out in the context of a full-cohort testing program in Western Australia, in which students aged

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

approximately 7, 9, and 11 years took part in reading, writing, numeracy, and spelling tests. Reading and numeracy were assessed using multiple-choice and short-response questions. To assess writing, students were required to write a narrative and their performances were assessed using a rubric. The same task and rubric were used across year levels.

The Nature of the Construct Assessed and Process for Developing the Rubrics

During the time the research was conducted, Western Australia had an Outcomes and Standards Framework. At that time, it was stated in the preamble that the frameworks articulate "typical learning achievements. They are a 'progress map' that describes how key concepts and skills develop as students achieve the learning outcomes set out in the Curriculum Framework" (Curriculum Council, 1998, p. 1). The framework divided the content area of the curriculum into eight learning areas. The English learning area comprised reading, writing, and speaking and listening. Each of these consisted of three strands that were elaborated using pointers to make concrete the specific learning outcomes in the content area (Andrich, 2002).

The two rubrics used in the empirical research were designed to assess aspects of the writing that fell into two categories. These may be described as follows: (i) authorial choices, which encompasses features of writing where the writer is free to make choices including subject matter, language choices, development of tone, style, voice and reader-writer relationship, and (ii) conventions, where the writer is expected to largely follow rules, including spelling, punctuation, correct sentence formation, and clarity of referencing. The development of the framework involved extensive consultation and expert input and drew upon a large number of student work samples. The framework determined the criteria in the narrative writing rubric. The criteria were therefore closely linked to the outcomes taught to students. The descriptions of ordered categories within each criterion were derived from writing outcomes in the framework. Work samples were used to exemplify each category for each criterion in the rubric, consistent with Wilson's (2005) approach.

Table 2 lists the nine criteria used to assess students' narrative writing. Both rubrics were trialed and refined before they were used in the state testing program. For a more comprehensive description of the conceptualization of writing captured in the rubrics, refer to the Australian National Assessment Program—Literacy and Numeracy (NAPLAN) writing rubric (Australian Curriculum, Assessment and Reporting Authority, 2010), which is available online. The NAPLAN rubric was based on the new rubric referred to in Stage 2 of the studies.

Overview

In the first stage of the empirical research, we observed a multimodal distribution of raw scores from the writing assessment in which a large proportion of students' scores were clustered on a relatively small proportion of the score points available. Because there was no multimodal distribution for reading or mathematics for the same population of students in the same assessment program, the distribution for writing appeared to be anomalous. The modes did not correspond with the means of the year groups. No other evidence accounted for the existence of a tri-modal distribution in writing. The most parsimonious hypothesis of the cause of the multimodal distribution was a halo effect resulting in a predominance of score patterns such as (1, 1, 1, …), (2, 2, 2, …) across the criteria. There were nine criteria and the modes corresponded with multiples of 9 (i.e., 9, 18, 27, etc.).

Where such score patterns predominate, there is an extreme lack of variation of ratings across criteria. It is therefore unlikely that for each separate criterion "variations in the attribute causally produce variations in the outcomes of the measurement procedure" (Borsboom et al., 2004, p. 1061). That is, it is unlikely that individual differences in separate aspects of writing performances are reflected in corresponding variations in the ratings on separate assessment criteria.

Preliminary analyses indicated that, as we had hypothesized, the multimodal distribution was caused by pronounced rating tendencies. To confirm this, separate technical research demonstrated that a multimodal distribution is produced where (i) there are multiple items each with the same number of performance categories and (ii) if a score of x is awarded on any one criterion, a score of x will subsequently be awarded to students in a larger ability range than would be the case in the absence of the halo effect (Andrich, Humphry, & Marais, 2012; Marais & Andrich, 2008, 2011). Such response tendencies constitute a specific form of violation of the assumption of local independence in item response models. Item response models are used in large-scale testing programs internationally.

Having confirmed that rating tendencies created the multimodal distribution, we turned our attention to the source. We investigated the source of the rating tendencies in several stages. These stages successively revealed that the cause of the rating tendencies did not correspond with one of the standard categories of definition of a halo effect given by Fisicaro and Lance (1990). We then developed the conceptual framework described above and investigated whether the design features of the rubric induced the apparent halo effect, as depicted in Figure 1. We did this by altering the structure of the rubric and testing whether there was evidence that the apparent halo effect had been reduced.

Methodological Approach

The methodological approach used in our series of research studies was an iterative tryout-redesign-tryout approach, involving "multiple approaches and modes of inquiry" (Ercikan & Roth, 2006). This approach involves multiple stages, and in this article, we focus primarily on the first and last of the stages. However, we also briefly describe interim stages, which were necessary for developing the conceptual framework and provided evidence that substantiates conclusions drawn from the research (see Figure 2).

The first and last stages of the research focus on (i) the original rubric used to assess writing in the state testing program and (ii) the reconceptualized rubric used in a subsequent year of the same testing program. For ease of reference, these are referred to as Stage 1 and Stage 2. Contrasting Stages 1 and 2 allows us to highlight the evidence showing that modifications to the rubric's structure and design resolved the rating tendencies.

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

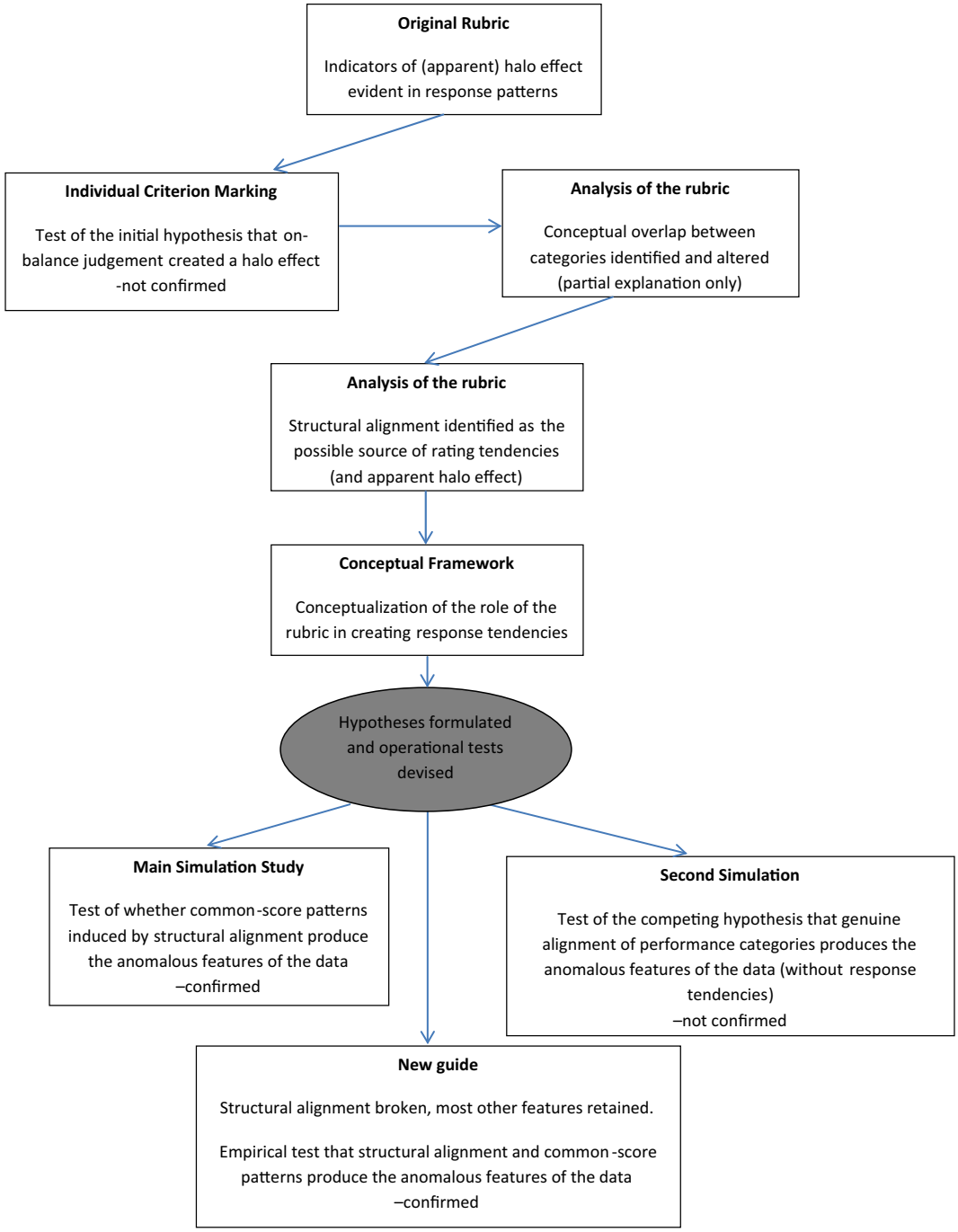

FIGURE 2. *Overview of the sequence of the research with the rationale for each component*

There were two interim or additional stages of the iterative tryout-redesign-tryout approach. In the first interim stage, we designed and implemented an individual criterion marking study to test the initial hypothesis that an on-balance (global) judgment by raters influenced ratings on other criteria to induce a halo effect. In this blinded experiment, each participant rated one criterion, such that a total of nine raters assessed any given student's performance. Contrary to our expectations, this experiment showed the halo effect *appeared* to exist even when raters did not make the on-balance judgment and each rater had no knowledge of the scores given by another maker for any other criterion. This ruled out the influence of a rater's general impression of a performance as the primary source of the rating tendencies. We then turned our attention to the content and structure of the rubric.

In the second interim stage, we analyzed the rubric and identified semantic and conceptual overlap among descriptions of gradations in separate criteria. It was evident that the descriptions of gradations in certain criteria closely reflected the descriptions of gradations in other criteria, which indicated conceptual overlap as shown in Table 1. We broke the conceptual overlap by changing the descriptions such that they shared no direct commonality. However, we preserved the a priori alignment of the gradations of quality so that most criteria still had the same

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

Table 2
Structure of Stage 1 Rubric for Writing Assessment

Criterion	Score Range	Criterion	Score Range
On-balance judgement* (OBJ)	0–7	Form of Writing (F)	0–7
Spelling (Sp)	0–5	Subject Matter (SM)	0–7
Vocabulary (V)	0–7	Text Organisation (TO)	0–7
Sentence Control (SC)	0–7	Purpose and Audience (PA)	0–7
Punctuation (P)	0–6		
		Total score range	0–60

*OBJ is a global judgment that acknowledges raters may judge that the "whole is greater than the sum of the parts."

Table 3
Participants in Each Study

	Number of Raters	Total Number of Writing Samples
Stage 1: Original rubric (full cohort testing program)	~200 per calendar year	~87,000 per calendar year
Individual criterion marking	27	632
Conceptual overlap trial	20	300
Stage 2: Reconceptualized rubric (full cohort testing program)	~200 per calendar year	~72,000 per calendar year

numbers of score points. We found that breaking the conceptual overlap among descriptions resulted in only a modest reduction in the apparent halo effect.

We then hypothesized that the a priori alignment of gradations of quality among criteria was the primary source of the apparent halo effect and that the conceptual overlap was a subsidiary source. It was therefore deemed necessary to "break" the a priori alignment of the gradations of quality; that is, we altered the rubric to avoid the same number of score points for each criterion. This led to the redesigned rubric that was used in Stage 2 of the empirical research.

The number of raters and total number of narrative performances for each stage of the tryout-redesign-tryout approach are shown in Table 3.

Procedures

Stage 1

In Stage 1, the rubric consisted of nine criteria. A narrative writing task was administered to approximately 87,000 students in total in Grades 3, 5, 7, and 9. The students in Grades 3 to 7 comprised the full cohorts in the Western Australian school system as part of the large-scale Western Australian Literacy and Numeracy Assessment (WALNA) program. Raters were required to make an on-balance judgment about the level of each student's performance and then were required to assess each performance in terms of the nine criteria shown in Table 2. Raters participated in extensive training, as detailed in the supplementary materials (available on the journal website).

The category descriptors for each criterion were derived directly from descriptions of the English learning area in the Western Australian Outcomes and Standard Framework (Curriculum Council, 1998). These described the typical progress students were expected to make from the commencement of school until high school graduation.

Students' writing performances were assessed using a rubric that had been deliberately constructed so that most of its criteria had common score ranges, with categories having been derived from generic descriptions of performance levels.

Stage 2

In Stage 2, a narrative writing task was administered to a total of approximately 72,000 students in total in Grades 3, 5, and 7 in a subsequent calendar year of the large-scale WALNA program. These students comprised the full populations in the relevant grades within the Western Australian schooling system.

A team responsible for English assessments in the large-scale program developed the new rubric based on a qualitative analysis of approximately 100 exemplars. The team compared the exemplars to identify and articulate observed, qualitative differences. During this process, there was no preconceived notion of the number of observable, qualitative differences for each criterion. Consequently, there was no reason that there should be the same number of gradations of quality for each criterion. Instead, the number of categories varied depending on the number of discernible qualitative differences, as shown in Table 4.

The new rubric no longer had a matrix structure. For example, in *vocabulary* and *sentence structure*, there were seven categories because in a representative range of student performances from Years 3 to 7, seven qualitative differences could be discerned and described. In *paragraphing*, however, only three qualitative differences could be distinguished, so there were only three categories.

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

Table 4
Revised Rubric for of Writing Assessment Used in Stage 2

Criterion	Score Range	Criterion	Score Range
On-balance judgement (OBJ)	0–6	Punctuation within sentences (PI)	0–3
Spelling (Sp)	0–5	Narrative structure (NS)	0–4
Vocabulary (V)	0–6	Paragraphing (Para)	0–2
Sentence structure (SS)	0–6	Characterisation and setting (CS)	0–3
Punctuation of sentences (PO)	0–2	Ideas (I)	0–5
		Total score range	0–42

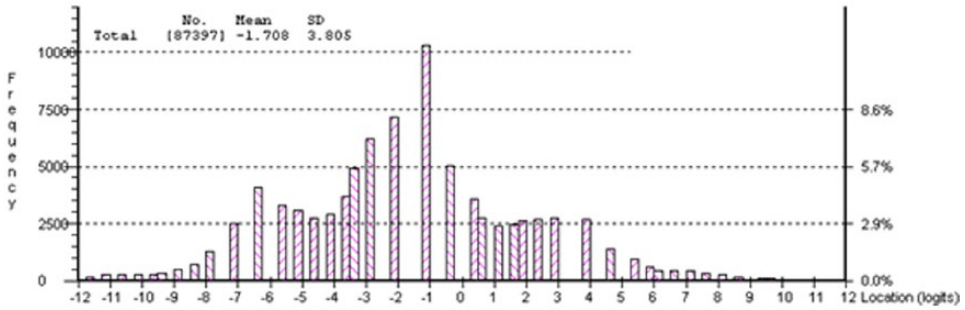

FIGURE 3. *Distribution of ability estimates for the original rubric used in Stage 1*

In using the new rubric, approximately 200 raters participated in extensive training as detailed in the supplementary materials (available on the journal website).

Empirical Results

Stage 1

In Stage 1, response data from the original and redesigned rubrics for writing were analyzed using the polytomous Rasch model (Andrich, 1978; Masters, 1982; Rasch, 1961; Wright & Masters, 1982). This model is used in the Australian context and in a range of large-scale testing programs internationally such as the Program for International Student Assessment (PISA).

Figure 3 shows the multimodal distribution of ability estimates obtained from the original rubric for students in all age groups in the testing program. The ability estimates were obtained by application of the polytomous Rasch model. Because there is a one-to-one correspondence between raw scores and ability estimates, the figure effectively shows the distribution of students' total scores. Approximately 55% of students have no more than two variations from a given score across nine criteria on the rubric. Consequently, three modes occur at or near total scores of 9, 18, and 27, corresponding with score vectors (1, 1, 1, …, 1), (2, 2, 2, …, 2), and (3, 3, 3, ..., 3). These total scores correspond directly with the scale scores at which the three modes occur at approximately -6.4, -1, and +4 in Figure 3.

Thus, a large proportion of total scores cluster on a relatively small proportion of the available score points because a high proportion of rating patterns fall on a very small proportion of all possible response patterns. Effectively, the whole population of students is categorized into three broad groups, which implies the rubric captures only coarse-grained information about student performance. An example of the consequence for teachers is that in one large high school, approximately 40% of students in a single calendar year fell within a range of just five score points (25–29).

Stage 2

Figure 4 shows the distribution of ability estimates obtained from the restructured rubric for students in all age groups in the testing program. It can be seen in Figure 4 that there is no longer a multimodal distribution and therefore that there is not the same tendency to categorize the whole distribution into just three broad performance groups.

Results are shown in Figures 3 and 4 for the same population of students in different calendar years. The same multimodal distribution was observed for 5 successive years prior to restructuring the rubric, whereas subsequently the smooth distribution with a single mode was observed for 3 successive years. The research effectively used an *equivalent-groups* design in which full populations of students in separate calendar years constitute the equivalent groups. (It was unnecessary to take random samples as full population data were available.) Because the change in the distribution of the full population occurred after the rubric was changed and the distribution and its features remained the same after that point, it is clear that the differences between the distributions result from differences between the designs of the rubrics rather than differences between the actual distributions of students on the trait.

The person separation index (index of internal reliability) for the original rubric was approximately 0.96. The person separation index for the new rubric was approximately 0.94. These

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

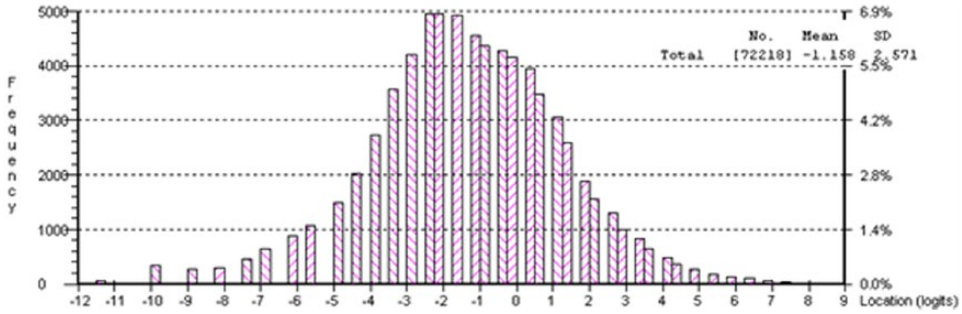

FIGURE 4. *Distribution of ability estimates for the new rubric used in Stage 2*

high levels of internal reliability were consistent over time, and therefore, the focus of the various stages of the research was validity and not reliability. It should be noted, however, that rating tendencies, like the halo effect, artificially inflate the (apparent) reliability and so values do not imply that the original rubric was more reliable (Bechger, Maris, & Hsiao, 2010).

Discussion

Next, we discuss the results from the empirical studies and the results of a related but separate study, and then we touch on the implications of the results with a particular focus on the conceptual framework.

Discussion of Results

As evident in Table 4, the score ranges in the restructured rubric have a different number of gradations of quality for each criterion. Removing the a priori alignment of categories removed the constraint on judgments that led to the predominance of specific response patterns and the resulting multimodal distribution (as seen in Figure 2). Using the restructured rubric, the raters were able to make independent judgments for each criterion. As a result, the distribution for writing had a single mode like the distributions of the same students for reading and mathematics.

Qualitative feedback from raters, in combination with detailed examination of scoring patterns across the criteria, indicated that by making independent judgments on each criterion, raters were able to discriminate between student performances in a more fine-grained way across the range of performance. As a result, the scores obtained from the reconceptualized rubric are more continuously distributed across the range of available score points (Figure 3). The importance of the continuous and smooth distribution of scores is that there is continuous variation, which can reflect fine-grained variation in the level of the construct on multiple aspects of the construct. That is, the variation in the total scores can more faithfully capture construct-relevant variance as necessary to meet the validity criteria articulated by Messick (1989) and Borsboom et al. (2004). We must stress that it is irrelevant whether the distribution is normal (bell-shaped). The key is that the distribution is smooth and that there is no longer a large proportion of students clustered on a very small subset of the score points; that is, the population is no longer effectively crudely categorized into a small number of groups of students.

In a separate study, Heldsinger and Humphry (2010) demonstrated the concurrent validity of the reconceptualized rubric used in Stage 2 of the research above, following the suggestion made by Reddy and Andrade (2010) to establish the "correlation between rubric-referenced scores and other measures of performance" (p. 446). In this separate study, a set of performances was marked with the rubric and assessed using an entirely different method of assessment, in the form of pairwise comparisons. Twenty judges each compared the quality of approximately 100 pairs of writing samples in order to rank them and to obtain scale scores. The study by Heldsinger and Humphry (2010) obtained a correlation of $r = 0.921$ between scale scores obtained from the two methods, indicating a high level of *concurrent validity* (the degree of agreement between results from two tests designed to assess the same construct).

In this article, we have reported an obviously trimodal distribution. We have observed similar distributions in writing results for a number of separate Australian state testing programs, as well as in school-level data derived from teacher judgments in Western Australia. The rubrics used in these contexts also had artificially aligned gradations of quality across criteria. It is noted that in these contexts, we did not always see distributions with pronounced modes. In all cases, we observed a predominance of scores associated with vectors such as (1, 1, 1, …), (2, 2, 2, …). In some cases, there were sufficient score patterns such as (1, 2, 1, 1, 2, …) to prevent the distribution from being multimodal, but the distributions still departed markedly from a normal distribution, often being strongly platykurtic (plateau-shaped). The shape of the distribution is merely symptomatic: The key problematic feature is the high occurrence of rating tendencies in the data collected using a rubric.

Implications

A direct implication of this research was that the reconceptualized rubric was adopted when Australia moved from state testing to a single Australian testing program, NAPLAN. The findings also constituted the central body of evidence instrumental in changing the way that teachers are required to assess student performance in both primary and secondary school in Western Australia (Andrich, 2006). The conceptual framework is intended to serve a basis for more general reconceptualization of the way in which rubrics are designed to optimize construct-relevant variance and construct representation.

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

Summary and Conclusion

We began this article by noting a lack of attention in the literature to Messick's challenge for evidence that rubrics used in scoring of performances capture fully functioning complex skills. We later noted there is a lack of attention in the literature to the potential role of the structure of rubrics to influence ratings in such a manner as to produce rating tendencies that may be mistaken for a halo effect.

In the series of studies comprising the tryout-redesign-tryout approach, we first conducted a blinded experiment to test the hypothesis that an initial on-balance judgment induced an apparent halo effect. It was established that the issue was not the raters' inability to treat each criterion independently but that the rubric itself forced judgments to be dependent, resulting in an apparent halo effect. Second, we conducted both qualitative and quantitative analyses to investigate whether removing semantic and conceptual overlap between criteria removed the apparent halo effect. These changes failed to remove the apparent halo effect, which indicated that the a priori alignment of gradations of quality among criteria was the most plausible source of the underlying rating tendencies. Once we had described the development of writing ability in terms of each criterion separately, the pronounced rating tendencies no longer existed. In the last stage, we redesigned the rubric to more faithfully capture qualitative gradations separately for each criterion. This resulted in a rubric with improved construct representation that allowed the assessments to better capture construct-relevant variance associated with each aspect of writing. A separate study established strong concurrent validity of the rubric, referenced to an entirely different method of assessment in the form of pairwise comparisons.

We have not attempted to demonstrate the generalizability of the findings presented in the research and recognize this limitation of the research to date. Although further empirical evidence will be invaluable, as we have stressed, the onus remains on those who develop and use rubrics to substantiate their validity (Messick, 1994; Reddy & Andrade, 2010). A critical next step is to investigate whether there is evidence of rating tendencies in other data sets obtained from rubrics in other contexts. Irrespective of whether such evidence is found, there is a pressing need to reconsider the prevalent matrix design of rubrics. In doing so, there is a fundamental need to consider whether there is any theory or empirical evidence to justify designing rubrics that comprise the same number of performance categories for multiple aspects of a construct.

NOTE

The work was supported by an Australian Research Council Linkage grant with the Australian Curriculum and Reporting Authority and Western Australian Curriculum Council as Industry Partners, on which Stephen Humphry and David Andrich are chief investigators. FundRef Funding Sources Australian Research Council (Grant/Award LP110100590).

REFERENCES

Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, *57*(5), 13–18.

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching*, *53*(1), 27–30.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 357–374.

Andrich, D. (2002). A framework relating outcomes based education and the taxonomy of educational objectives. *Studies in Educational Evaluation*, *28*, 35–59.

Andrich, D. (2006). *A report to the Curriculum Council of Western Australia regarding assessment for tertiary selection*. Curriculum Council, Western Australia, Australia. Retrieved June 19, 2012, from http://platowa.com/documents/Andrich/report.pdf

Andrich, D., Humphry, S., & Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement*, *36*(4), 309–324.

Australian Curriculum, Assessment and Reporting Authority. (2010). *NAPLAN 2010 writing narrative marking guide*. Retrieved March 5, 2014, from http://www.nap.edu.au/verve/_resources/2010_Marking_Guide.pdf

Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement*, *34*, 607–619.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Bresciani, M. J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., & Hickmott, J. (2009). Examining design and inter-rater reliability of a rubric measuring research quality across multiple disciplines. *Practical Assessment, Research and Evaluation*, *14*(12). Available at http://pareonline.net/

Curriculum Council. (1998). *Outcomes and standards framework: Overview, student outcomes statements*. Osborne Park, Australia: Curriculum Council of Western Australia.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, *64*(1), 5–30.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, *27*(2), 14–24.

Ercikan, K., & Roth, W.- M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher*, *35*(5), 14–23.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, *14*, 419–429.

Fisicaro, S. A., & Vance, R. J. (1994). Comments on the measurement of halo. *Educational and Psychological Measurement*, *2*, 366–371.

Heldsinger, S., & Humphry, S. M. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–20.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130–144.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, *65*, 507–516.

Kohn, A. (2006). The trouble with rubrics. *English Journal*, *95*(4), 12–15.

Marais, I., & Andrich, D. (2008). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*, 200–215.

Marais, I., & Andrich, D. (2011). Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, *12*(3), 194–211.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Meier, S. L., Rich, B. S., & Cady, J. (2006). Teachers' use of rubrics to score non-traditional tasks: Factors related to discrepancies in

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014

scoring. *Assessment in Education: Principles, Policy and Practice*, *13*(1), 69–95.
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.
Messick, S. (1994). The interplay of evidence and consequences in validation of performance assessment. *Educational Researcher*, *23*(2), 13–23.
Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10).
Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460–515). Maple Grove, MN: JAM Press.
Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–143.
Perlman, C. C. (2003). Performance assessment: Designing appropriate performance tasks and scoring rubrics. *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators*. Retrieved June 13, 2012, from Education Resources Information Centre.
Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership*, *55*(2), 72–75.
Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV (pp. 321–334). Berkeley: University of California Press.
Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment and Evaluation in Higher Education*, *35*(4), 435–448.
Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39.
Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, *34*, 159–179.
Silvestri, L., & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education*, *9*, 25–30.
Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 25–29.
Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research and Evaluation*, *9*(2), 2004.
Wald, H. S., Borkan, J. M., Scott Taylor, J., Anthony, D., & Shmuel, P. R. (2012). Fostering and evaluating reflective capacity in medical education: Developing the REFLECT rubric for assessing reflective writing. *Academic Medicine*, *87*(1), 41–50.
Wiliam, D. (2011). *Embedded formative assessment*. Bloomington. IN: Solution Tree.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Portsmouth, NH: Heinemann.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

AUTHORS

STEPHEN MARK HUMPHRY, PhD, is an associate professor at the Graduate School of Education, The University of Western Australia, M428 35 Stirling Hwy Crawley WA 6009 Australia; *stephen.humphry@uwa.edu.au*. His research focuses on both conceptual and applied issues in educational measurement, and he has extensive experience in large-scale assessment in Australia through his industry work and research grants.

SANDRA ALLISON HELDSINGER, EdD, is a senior research associate at the Graduate School of Education, The University of Western Australia, M428 35 Stirling Hwy Crawley WA 6009 Australia; *sandy@assessmentcommunity.com.au*. Her research focuses on conceptual and applied issues in educational assessment, and she has extensive experience in large-scale and school-based assessment.

Manuscript received September 19, 2012
Revisions received April 22, 2013, and March 25, 2014
Accepted June 8, 2014

Downloaded from http://er.aera.net at East Carolina University on October 9, 2014